BALANCING MULTIPLE DEMANDS: SAMPLE REDESIGNS FOR THE NSF SYSTEM

Linda Hardy, National Science Foundation

KEY WORDS: Demographic surveys, multiple sampling goals

One of the responsibilities of the National Science Foundation (NSF) is to collect, organize and analyze data relating to Science and Engineering. Within the Foundation, much of this work is done by the Division of Science Resource Studies, which sponsors a variety of data collection efforts. The Scientific and Technical Personnel Data System (STPDS) is a system of data about scientists and engineers (S&E) and is developed from three of the Division's surveys: the National Survey of College Graduates; the Survey of Recent College Graduates; and the Survey of Doctorate Recipients. These are surveys of individuals and have cross-sectional and longitudinal components. In 1990, NSF began a broad-based effort to redesign and unify the STPDS surveys with the goal of implementing the changes coincident with the large baseline survey of scientists and engineers that is done after each decennial census. This paper discusses the work that has been done to develop integrated sampling goals and principles for the STPDS system and to overcome the sampling related problems of the past.

Universe Definition:

In the late 1980s, NSF asked the National Research Council's Committee on National Statistics to review and recommend improvements in the STPDS. One of CNSTAT's recommendations urged the creation of a new unified and conceptually consistent definition of scientists and engineers—a definition that was, to the extent possible, also consistent with standardized data definitions used by other government agencies.

In the STPDS redesign, a simplified S&E universe definition was developed. The prior S&E definition attempted to combine an individual's education and occupation into a single characteristic. Complex decision criteria were needed to determine if the individual was in the scope of the survey. These complexities also caused less stability in the classification of sampled individuals over time. The new definition treats education and occupation as separate characteristics and defines S&E as anyone possessing either. No attempt is made to create a combined variable. Maintaining the separation is more consistent with uses of the data such as analyses of supply/demand and of educational outcomes. A side benefit to the new definition is a somewhat more stable S&E population size. Factors contributing to stability are: education is a permanent characteristic; more than 90% of the S&E target population have at least one S&E degree; and the reporting of educational degrees is generally good and subject to less response variance.

The STPDS universe definition of S&E is: (1) Persons holding a Bachelor's degree or higher, with at least one degree in an S&E field, and/or (2) Persons holding a bachelor's degree or higher in any field and working in an S&E occupation, as of the survey reference date. The target population is limited to non-institutionalized persons, *75* years old or less, who are residents of the United States or U.S. Territories on the survey reference date.

As can be expected, the STPDS definition reflects practical limitations in sample frame coverages, data and budget. For example, persons with S&E technology training, at less than the bachelor's level, are excluded from the S&E definition despite the fact they are of interest to economic studies of S&E issues.

Despite the limitations, the new STPDS definition is broader than the previous definition and is expected to include about 10-12 million persons.

STPDS Sample Frames and Coverage:

The National Survey of College Graduates (NSCG) provides baseline estimates for most of the STPDS target population. It represents the "stock" of the bachelor's and master's degreed S&E population at the beginning of the decade. (the survey also provides estimates for the a small part of the PhD population—those who received their degrees abroad. After the initial survey the NSCG is subsampled to provide a panel of experienced S&E which are followed throughout the decade.

The NSCG frame consists of individual respondents to the 1990 decennial census sample of households (the long form)—those who indicated their highest degree was a bachelor's, master's or professional degree, approximately 38 million records. The frame is somewhat inefficient for the STPDS S&E definition, however, because the field of degree was not captured. The census long form sampling rates for households were 1/2, 1/6, and 1/8, with all individuals surveyed in a selected household. A series of adjustments were made to the person record weights after data collection. Principally, controlling the weights to the full census counts at detailed demographic and geographic levels.

Survey of Doctorate Recipients (SDR) provides estimates for the PhD population that received their S&E doctorate from a US institution. The sample frame is based on records collected by the Survey of Earned Doctorates (SED). The SED questionnaire, a self administered form, is collected by the university from persons to whom they have awarded a PhD. The university provides the information to the National Research Council which maintains the longitudinal Doctorate Records File (DRF)—the frame used by the SDR. Institutional and individual participation is very high with DRF frame coverage of those receiving PhDs from US institutions estimated at greater than 99 percent. The excellent coverage and degree and demographic characteristics information available on the DRF frame allows for efficient oversampling of S&E subpopulations of special interest. The basic design is a stratified simple random sample with unequal weights resulting from oversampling.

The Survey of Recent Science and Engineering Graduates, also called the New Entrants Survey, provides estimates for individuals receiving S&E degrees at the bachelor's and master's levels. The survey represents the flow of individuals into the "educated as S&E" STPDS target population. Bachelor's and master's S&E degrees represent about 95 percent of the number of S&E degrees granted annually. These data on the flow of degree recipients data cannot be provided by the NSCG survey because its frame is updated only once a decade. Over the decade, the *proportion* of the total STPDS target population represented New Entrant's records grows while that represented by the NSCG panel declines. A frame for the New Entrants Survey presents special problems since no central roster of graduates exists, as with the SDR. The New Entrant frame must be developed in a two stage process. The first stage is a sample of degree granting schools from the frame of accredited institutions (the Department of Education's WEDS file, Integrated Postsecondary Education Data System. WEDS has virtually 100 percent coverage of degree granting schools. Typically, the first stage selection probabilities are proportional to the number and types of S&E degrees awarded by the school. Increased probability is given to schools that have proportionally higher numbers of women and minorities in order to enhance the ability to oversample these subgroups later. The second stage is to develop the frame of individuals by obtaining lists of graduates by degree field along with relevant data collection information.

STPDS Coverage: In general, the coverage of the STPDS surveys is good with most of the stock and flow of the S&E target population covered by the three survey sampling frames. There are coverage problems, however. One is persons who work in S&E occupations but do not have S&E degrees. Information about this

subpopulation is collected in the first NSCG but not in its subsequent panel surveys. The New Entrant and the SDR do not provide information about this subpopulation because their frames are of individuals who have earned S&E degrees. A second problem is coverage of persons holding S&E degrees from non-US institutions. Most of these are immigrants who have entered the US after having completed their education. As before, an estimate for this subpopulation can be made from the first NSCG but subsequent information is not available. There is some duplication in coverage between the STPDS survey frames. These include persons receiving a second degree in an S&E field after census day, April 1990, and foreign-born PhDs who hold S&E degrees from US institutions.

In addition to coverage errors, other data gaps in the STPDS system result from the frequency of the surveys—every two years—and the transient nature of some variables, such as occupation and disability status. In concept, the system can provide cross-sectional, time series and longitudinal data, but this is limited by the frequency of the surveys. To improve the availability of data in the STPDS, NSF will be investigating the use of supplemental data from other sources and the applicability of statistical modeling methods.

Sample Design Goals and Issues:

There are two goals for the sample redesigns of the STPDS surveys. The first is recognition of the wide diversity of the uses and users of the STPDS data. The second is allowance for S&E subpopulations of special interest, such as women and minorities. For example, NSF is mandated to produce a report to Congress on the status of women and minorities in science and engineering every two years. The principal task of the redesign work is to develop sample designs that realistically balance these two goals. Implicit in this work is viewing the STPDS as an integrated system--although the differing sampling frames mean that the principles developed here are guidelines rather than strict criteria.

User Diversity: The variety of analyses and broad interests of the STPDS data users have important effects on the sample redesigns for the surveys. Sample designs typically optimize on one or a limited subset of the variables. The choices made are based on user priorities and policy requirements. Unfortunately, it is not always possible (or desirable) to rationally order these elements and assign individual precision level targets—a basis for many sample designs. Using a few variables as proxies or building a compound variable from a wide array may poorly serve the broader design goals. Given fixed sample sizes, emphasizing one set of variables or group of users requires sample loss in other areas. Unlike other large demographic surveys, STPDS users place greater emphasis on measuring a characteristic of a subpopulation rather than estimating the subpopulation size itself. Estimates for the S&E population as a whole are often of lesser interest. Users also tend to focus on one or a limited number of educational or occupational fields rather than a range of fields.

With sample/budget constraints, user diversity would argue for simple proportional samples—modified as needed to adjust for problems with the different sampling frames. With proportional allocation, increased precision is coincident with the size of the population. (Target CVs are not chosen for each strata.) Proportional allocation, which has equal weighting of the sample members, maintains flexibility for future research, the scope of which is difficult to predict in advance, and for micro data users who create their own variables and analysis cells. Because of these advantages, a proportional sample is being used as the starting point for the redesign of the survey samples.

One simplification that has been used is to compare the various candidate sample designs using a fixed proportion ($p = 0.2$), thus fixing the strata variances. This approach was preferred because it allowed direct comparisons of the relative effects of design changes on the sample allocation, weighting and CVs. Earlier sample designs, in the 1980s, used the proportion of persons working in business and industry as the

design variable. This variable was intended to be a proxy for the many others but at times created arbitrary differences in strata variances.

Special Subpopulations: While we felt it was not appropriate to specify target precision levels for each strata within the STPDS, it was important to recognize the need to increase the reliability of data for some subgroups—particularly women and minorities. This was done by creating a large number of strata and setting a minimum sample size and/or CV for these cells. The goal was to develop an general oversampling scheme that would recognize these subpopulations without letting the oversampling dominating the designs.

Common sampling strata were developed for the STPDS surveys. The strata used in the redesign reflect variables of the most interest to a broad group of data users. The intersection of four variables form the strata--occupational/educational fields, degree level, sex, and special demographic characteristics, the last is termed the "NSF group" variable. In general, the strata configuration was not chosen to isolate between strata variance, although some reduction in overall variance is expected to result from the occupational/educational field variable.

The last variable, NSF group, is a combined variable that was constructed to reduce the number of strata. In the early stages of the redesign it was recognized that an almost infinite number of strata could be created from demographic characteristics. The NSF group variable combines race/ethnicity, disability status, country of birth and citizenship into 8 groups: 1-US-born disabled persons group; *5* US-born non-disabled race/ethnicity groupings; foreign-born US citizens; and foreign-born non-US citizens.

The educational/occupational field variable is a "conceptual" variable in the STPDS system, since none of the survey frames contain both the education and occupation variable. To better unify the survey designs, similar stratification of the fields groupings is used on all the surveys. The SDR and New Entrants survey are based on frames that originate at the degree conferring institution and thus have accurate measures of educational field on the frame. The benefit is that, in the SDR and the New Entrant survey, virtually all the sampled individuals are in the target population. The NSCG frame contains information on the level of the degree and 1990 census occupational classification, but no information on the individual's field of degree is available. Unfortunately occupation is a difficult variable to collect accurately and is a transient characteristic. The lack of degree field information on the NSCG frame results in approximately half of the sampled persons being outside the scope of the STPDS definition. This large out-of-scope problem is the most significant technical problem faced by the NSCG.

The STPDS definition requires estimates for both education and occupation from each survey. Estimates for the "missing" field data are formed by aggregating sampling cells designed around the field characteristic available on the frame. However, the mapping of occupation to education (or the reverse) is not always good and this results in a great many sampling cells contributing to the estimate and, unfortunately, the "analysis cell's" variance.

Sample Allocation Guidelines:

Limiting Sampling Rates: The most important allocation criterion in the STPDS sample designs is a limitation placed on the differences in the sampling rates. This limitation is similar to informal criterion used by other agencies and is especially important to the STPDS surveys for a number of reasons.

As discussed earlier, data users are often more interested in estimates from small S&E subpopulations, than in higher level aggregate estimates. The subpopulation cells can be quite small in size and the data for them are often unreliable. Variance and bias effects may be large even when the sample is heavily supplemented. (The total S&E population as a whole is small.) Given the user interest in the small population cells, the natural tendency would be to reallocate much more sample to

them. Since they are so small, a little sample supplementation could be expected to have a significant impact—on variance at least. The problem with this reasoning relates to the variance effect discussed earlier, that is the sampling strata may not be the analysis cells that the users wish to analyze. For example, if a user wishes to analyze a characteristic, say percent of time physicists spend on research work, the estimate will need to be constructed from many sampling cells. The precision (CV) of the "analysis cell" will reflect the contribution to variance of all the cells that report persons having the occupation physicists. While "physicists" sampling cells will probably map well to the "physicist" analysis cell—justifying a heavy sampling rate—many other unlikely sampling cells will also contribute significantly to the variance. When the range (spread) in the rates is great, the analysis cells are more vulnerable.

In the STPDS sample designs of the 1980s, the sampling weights were allowed to vary as much as a thousand-to-one. Given the user interest in smaller cells, these types of designs proved to be very unwise. For the STPDS designs of the 1990s a limit on the range of rates has been set at 8 to 1 or less forming a central theme in the STPDS sample allocation processes.

Oversampling: Another allocation criterion is oversampling to meet a minimum sample size or minimum CV for the subpopulations of special interest. This criterion is constrained by the limitation on the range in rates discussed above. The purpose of setting minimum rather than target CVs is to provide an oversampling method that supplements rather than dominates the sample design. A minimum sample size was used in both the 1991 SDR sample design and the 1993 NSCG. This was based a criterion of 50 responses to a publishable cell. The NSCG also had some reallocation of sample to some race/ethnic groups to improve the CVs for major occupational groups.

Future Redesign Work:

By 1995 all the STPDS surveys will have samples redesigned to follow the guidelines outlined in this paper. While the differing sampling frames do not allow strict adherence to specific sample design criteria, the generalized strata and the principle of tightly controlling the range of sampling rates do provide consistency across the STPDS system. Oversampling using a minimum sample size provides support for the largest variety of data users, while reallocating sample to achieve a minimum CV improves the precision for minority and other subgroups of special policy interest. Finally, the use of consistent sample design guidelines across the STPDS surveys will provide data more useful to the STPDS system as a whole.